AD NUMBER

AD134787

LIMITATION CHANGES

TO:

Approved for public release; distribution is unlimited.

FROM:

Distribution authorized to U.S. Gov't. agencies only; Administrative/Operational Use; MAY 1957. Other requests shall be referred to Office of Naval Research, Arlington, VA 22203.

AUTHORITY

ONR ltr 26 Jun 1971

RELIABILITY FOR THE LAW OF COMPARATIVE JUDGMENT*

A Technical Report
prepared by

Harold Gulliksen
Princeton University
and
Educational Testing Service

and

John W. Tukey
Princeton University

May 1957

*Prepared in connection with research sponsored by the Office of Naval Research and the National Science Foundation

RELIABILITY FOR THE LAW OF COMPARATIVE JUDGMENT[1]

In studies using the method of paired comparisons and the law of comparative judgment, it is desirable to determine the reliability of the scales which are obtained. For a given set of data one might like to know the extent to which the law of comparative judgment is successful in accounting for the total variance in the data.

Mosteller (13) has outlined a chi-square test of the agreement between the fitted proportions (p^*) and the observed proportions (p); such a test labels the discrepancy between observation and theory as either "significant" or "non-significant" but does not indicate whether the variance accounted for by the theory is large or small in relation to the total variance in the data.

This property of significance tests is well known and has been clearly stated by Cochran (3) in his discussion of the chi-square test.

> "The power of the test to detect an underlying disagreement between theory and data is controlled largely by the size of the sample. With a small sample an alternative hypothesis which departs violently from the null hypothesis may still have a small probability of yielding a significant value of χ^2 . In a very large sample, small and unimportant departures from the null hypothesis are almost certain to be detected."

If the sample is small then the χ^2 test will show that the data are "not significantly different from" quite a wide range of very

---

[1]Thanks are due to Ledyard Tucker and Frederic Lord for valuable suggestions on the development presented here.

different theories, while if the sample is large, the χ^2 test will show that the data are "significantly" different from those expected on a given theory even though the difference may be so very slight as to be negligible or unimportant on other criteria. Fisher (6) gives a good illustration of this point in his analysis of Weldon's data on dice throws. If we test the theory that a throw of 5 or 6 has a probability of 1/3, then chi-square for Weldon's data is very large, with p of .0001. However, a very slight change in the theory -- from a probability of .3333 to a probability of .3377 -- gives a quite reasonable chi-square with a p value of .3 or .4.

In order to proceed appropriately in any scientific investigation it is likely to be necessary to answer two different questions:

1. Is it reasonable to say that random variation accounts for the difference between theory and data?
2. How large is this difference relative to the variation that is accounted for by the theory?

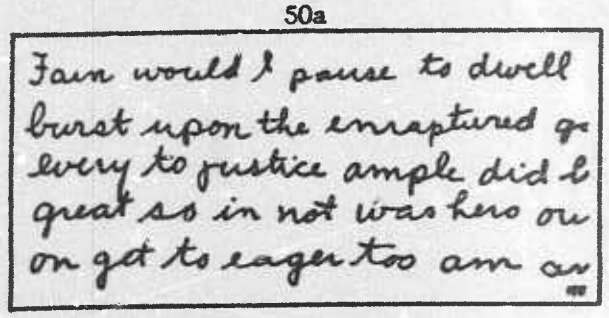
In studying the applicability of the law of comparative judgment, variance-component and analysis-of-variance techniques can provide appropriate answers to these questions by methods outlined below and there applied to two sets of data on handwriting specimens and to Mosteller's (13) baseball data.

The data of the example.

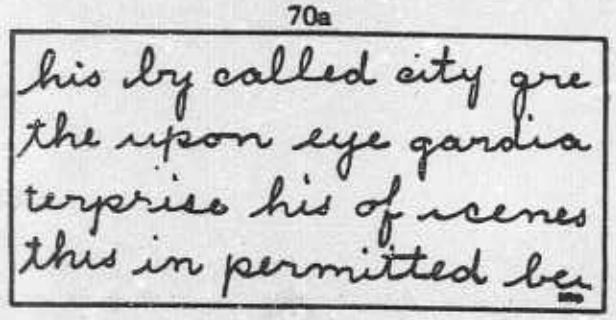
The handwriting specimens were chosen from the Ayres (1) handwriting scale. This scale consists of a series of handwriting specimens of nine different scale levels, numbered from 10 (the lowest) to 90

(the highest). Each of these scale values is represented by three specimens, a "vertical" style (a), a normal slant (b), and an extreme slant (c). Thus the scale consists of 27 different handwriting specimens. In conventional use, a handwriting specimen to be scaled is judged to be like one of the scale specimens or to fall between two of them. Thus, specimens can be scaled 10 to 90. The extremely bad or good ones might be either below 10 or above 90 respectively. Nine of these handwriting specimens were chosen for the present experiment: 50a, 50b, 50c, 70a, 70b, 70c, 80a, 80b, and 80c (shown in Figure 1). The 36 possible pairs for these nine specimens were arranged in a booklet, with instructions for the judge to pick the better member of each pair. It is interesting to note that one can easily develop a discussion in a class in measurement to indicate that there are numerous criteria on which it is possible to judge these handwriting specimens; the class will rather readily reach the conclusion that any set of judgments would be meaningless, highly unreliable, and unduplicatable unless one defined in great verbal detail exactly what characteristic was to be judged, instead of simply using the term "better handwriting." In the late 1930's this schedule was given without preliminary discussion of the problem to 100 students at the University of Chicago, and in the late '40's it was given, again without preliminary discussion, to 100 students at Princeton University. The data (p , the observed proportions,) are shown in Table 1. The agreement between these two sets of judgments for 100 people taken in different institutions about ten years apart is rather striking.

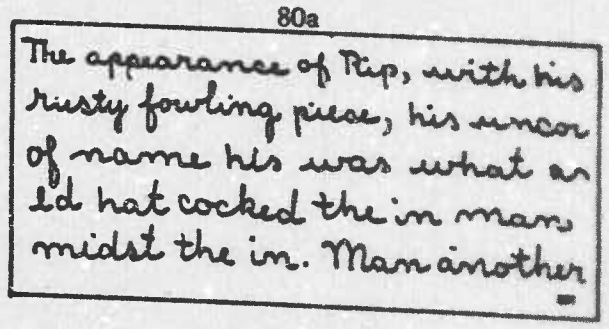
50a



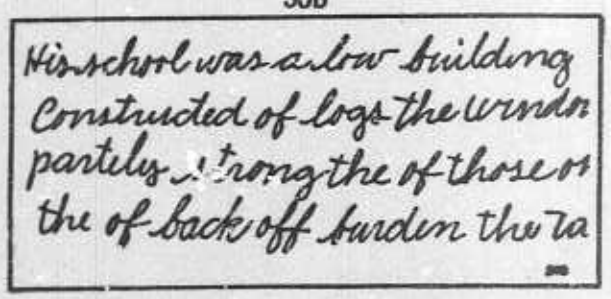
70a



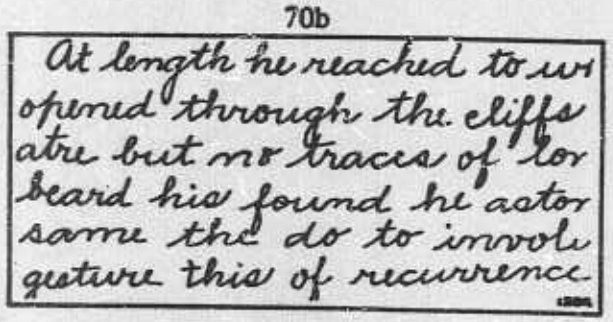
80a



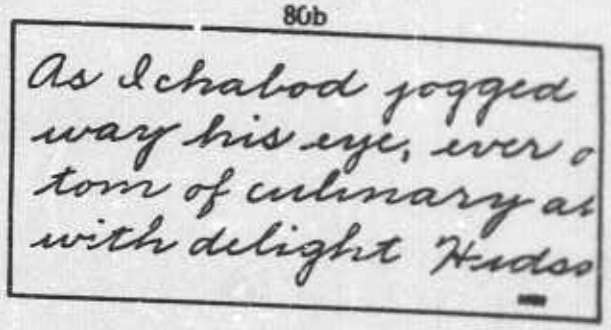
50b



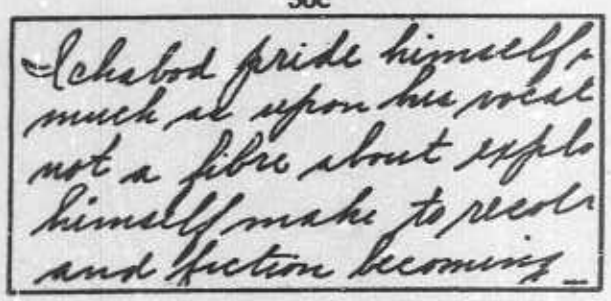
70b



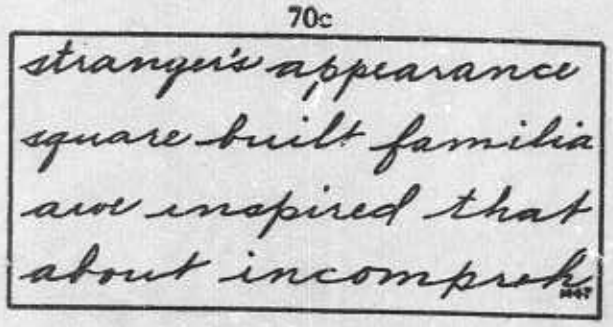
80b



50c



70c



80c

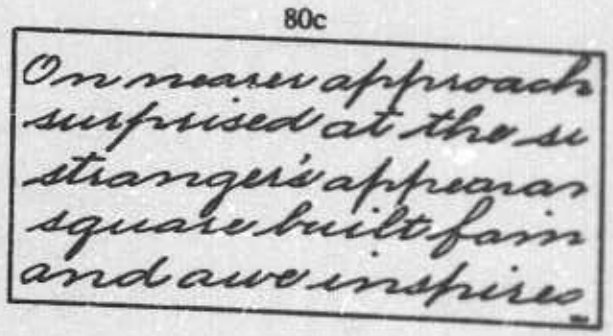


Figure 1. Selected Specimens from the Ayres Handwriting Scale.

The two sets of scale values obtained from utilizing the law of comparative judgment as stated by Thurstone (14, 15) are shown in Table 2. In both of these scales, stimulus 50a (the poorest one) has been chosen as having a scale value of zero. The fitted proportions (p*) computed from these scale values are given in Table 3. The scale values for the total group, given in Table 2, are found by summing the frequencies for the two groups and then proceeding to scale as for the single groups.

When Mosteller's (13) chi-square test for goodness of fit is applied to these data one finds (see Table 5, χ^2_D) a chi-square of about 74 for the Chicago data, 76 for the Princeton data, and 127 for the two groups combined. The corresponding p -values are each less than .0001, the chi-square value at the .01 level being only 48. Thus, the conclusion reached would be that the data are not fully accounted for by the law of comparative judgment. However, it is interesting and meaningful to know whether the fraction of the systematic variation which is not accounted for should be regarded as approximately 1 or 2 percent or as much as 75 percent. For example, if an aptitude test has a validity coefficient of .5 for predicting some criterion, it is considered a very useful test, even though it is also true that 75 percent of the variance in the criterion is not accounted for by the test. Under such circumstances it would doubtless be true that the criterion contains a significant non-random component that is different from anything represented by the test. Analysis-of-variance and variance-component analysis procedures will give information on the percentage of the variance which is accounted for and on the percentage which

TABLE 1

Experimental Proportions (p)

Handwriting Specimens →		50a	50b	50c	70a	70b	70c	80a	80b	80c
50a	C	--	.52	.67	.95	.99	.98	.99	.97	.94
	P	--	.52	.66	.88	.98	.98	.97	.83	.86
50b	C	.48	--	.60	.85	.95	.96	.98	.98	.95
	P	.48	--	.60	.69	.97	.96	.93	.94	.91
50c	C	.33	.40	--	.76	.78	.92	.91	.86	.96
	P	.34	.40	--	.70	.82	.94	.92	.84	.93
70a	C	.05	.15	.24	--	.76	.87	.95	.79	.78
	P	.12	.31	.30	--	.78	.84	.91	.70	.83
70b	C	.01	.05	.22	.24	--	.74	.80	.52	.71
	P	.02	.03	.18	.22	--	.64	.78	.37	.61
70c	C	.02	.04	.08	.13	.26	--	.59	.26	.56
	P	.02	.04	.06	.16	.36	--	.71	.30	.58
80a	C	.01	.02	.09	.05	.20	.41	--	.15	.31
	P	.03	.07	.08	.09	.22	.29	--	.15	.38
80b	C	.03	.02	.14	.21	.48	.74	.85	--	.61
	P	.17	.06	.16	.30	.63	.70	.85	--	.70
80c	C	.06	.05	.04	.22	.29	.44	.69	.39	--
	P	.14	.09	.07	.17	.39	.42	.62	.30	--

C = Chicago data

P = Princeton data

TABLE 2

Scale Values for Handwriting Specimens

	50a	50b	50c	70a	80b	70b	80c	70c	80a
Chicago	0.000	0.210	0.657	1.179	1.799	1.738	2.054	2.169	2.472
Princeton	0.000	0.107	0.384	0.808	1.252	1.578	1.690	1.794	2.048
Total Group	0.000	0.147	0.492	0.958	1.473	1.624	1.833	1.949	2.213

[Probability of choice approximately given by difference of scale values interpreted as a unit (standard) normal deviate, fitted according to Thurstone (14, 15) or Mosteller (13)]

TABLE 3

Theoretical Proportions (p*) Computed from

Scale Values in Table 2

		50a	50b	50c	70a	70b	70c	80a	80b	80c
50a	C	--	.583	.744	.881	.959	.985	.993	.964	.980
	P	--	.542	.650	.790	.943	.964	.980	.895	.955
50b	C	.417	--	.673	.834	.937	.975	.988	.944	.967
	P	.458	--	.609	.758	.929	.954	.974	.874	.943
50c	C	.256	.327	--	.699	.860	.935	.965	.873	.919
	P	.350	.391	--	.664	.884	.921	.952	.807	.904
70a	C	.119	.166	.301	--	.712	.839	.902	.732	.809
	P	.210	.242	.336	--	.780	.838	.893	.672	.811
70b	C	.041	.063	.140	.288	--	.667	.769	.524	.624
	P	.057	.071	.116	.220	--	.585	.681	.372	.545
70c	C	.015	.025	.065	.161	.333	--	.619	.356	.454
	P	.036	.046	.079	.162	.415	--	.600	.294	.459
80a	C	.007	.012	.035	.098	.231	.381	--	.251	.338
	P	.020	.026	.048	.107	.319	.400	--	.213	.360
80b	C	.036	.056	.127	.268	.476	.644	.749	--	.601
	P	.105	.126	.193	.328	.628	.706	.787	--	.669
80c	C	.020	.033	.081	.191	.376	.546	.662	.399	--
	P	.045	.057	.096	.189	.455	.541	.640	.331	--

C = Chicago data

P = Princeton data

remains to be accounted for after the law of comparative judgment has been utilized, and will thus give coefficients which are analogous to "reliabilities." For various illustrations of analysis of components of variance see, for example, Mood (12), Bennett and Franklin (2), Chapter 7, Davies' (4) discussion of "expectation of mean square" beginning in Chapter 4, Duncan (5), especially Chapters 23 and 24, or Tippett's (16) discussion of substantive variances in Chapters 6 and 7.

Framework of the analysis.

Since we are dealing with proportions, the sampling variance is a function of the true proportion as well as of the sample size. [$N\sigma_p^2 = \pi(1 - \pi)$] . If the analysis is conducted in terms of an angular transform of each proportion, then the (binomial) sampling variance is a function primarily of N , and not of the true proportion. The angular transform of the data is defined on different scales by different authors. The simplest scale for our purposes is that used by Hald (9) in his table, where

$\theta = 2 \text{ arc sin } \sqrt{p}$ (the arc is expressed in radians).

The variance of θ is $1/N$ approximately, for proportions not too near 1 or 0. If Np and $N(1 - p)$ both exceed 4 or 5 the approximation is quite good. Even more extreme cases may be analyzed by the use of the averaged angular transformation, Freeman and Tukey (8), which will be satisfactory for Np , $N(1 - p) \geq 1$. In the other common version, tabled by Fisher and Yates (7),

$\theta_F = \text{arc sin } \sqrt{p}$ (the arc is expressed in degrees).

The variance of θ_F is approximately 821/N for proportions not too close to 1 or 0. Thus if $p = .50$, $\theta = \pi/2 = 1.5708$, while $\theta_F = 45.00$. In general,

$$\theta_F = \frac{45.00}{1.5708}\theta = \theta\sqrt{821}$$

If tables of θ_F are used, then, in order to fit into the pattern of Table 4, the resulting sums of squares should be divided by 821.

The convenience of an analysis in terms of θ-values lies in the fact that for pure binomial variation the variance of any θ is substantially equal to the reciprocal of the number of observations on which the p is based. This property of the angular transformation allows the definition of modified chi-squares, such as the one used by Mosteller, which do not require denominators. When necessary, we shall distinguish these modified chi-squares as angular chi-squares.

For each ordered pair of stimuli (i,j) we have an observed angle θ corresponding to the observed p's of Table 1, and a fitted angle θ^* derived from the fitted scale and corresponding to the fitted p^*'s of Table 3. Because of the symmetry of the situation the mean of the complete set of p's, or that of the p^*'s, is .50. Correspondingly, the mean of any complete set of θ's and the mean of any complete set of θ^*'s equals 1.5708.

Using angles, the analysis of variance is given in terms of the following definitions:

$\theta = 2 \text{ arc sin } \sqrt{p}$ (observed values)

$\theta^* = 2 \text{ arc sin } \sqrt{p^*}$ (fitted values)

$\bar{\theta} = 1.5708 = 2 \text{ arc sin } \sqrt{.5}$

(the arc is measured in radians).

If all the stimuli are identical, and are judged to be identical, then the proportion of judgments " i greater than j " would be .5 in every case.

We treat the observed angles θ as if they were a sum of three types of contribution. This treatment is approximate in two ways. First, as Mosteller, (13, p. 213) was careful to point out in connection with his chi-square, the fitting used is a least-square fit on the normal scale but not on the angular scale. Consequently, residuals on the angular scale will not be as small as those resulting from a fitting procedure tailored to the angular scale. As a consequence, our estimated "reliability" coefficients will be somewhat smaller, just as Mosteller's chi-squares are somewhat larger, than those obtainable from more closely tailored fits. Second, the imperfect linearity of the relation of angles to normal deviates means that the true scale difference for any pair compared is, when measured in angles, only approximately a difference. For the purpose of defining variance components and reliabilities this latter effect should not be quantitatively important. We shall use these approximations freely, usually without further ado. (We hope to return to their consideration, as well as that of other refinements of procedure, in another paper.) Let us return to the three types of contribution associated with a single comparison (as of two specimens of handwriting) and contributing to the observed angle.

One contribution is approximately the difference between the <u>true scale values</u> for the two stimuli, (say $s_i - s_j$). These s values

may be thought of as drawn from a population with variance σ_s^2 . Hence the values in the cells ($s_i - s_j$) are regarded as drawn from a population with variance $2\sigma_s^2$.

Another is a <u>deviations component</u> (designated d) due to the deviations of the data from the linear scaling model used. These d -values are drawn from a population with variance σ_d^2 .

Due to the fact that we are dealing with values determined from proportions, we have a <u>binomial error</u> component (say b). These values are drawn from a population with variance σ_b^2 .

Thus we have the approximate composition of the observed values and the associated variance of the population from which each of these three quantities may be thought of as drawn, as follows:

$$\theta_{ij} \doteq (s_i - s_j) + d_{ij} + b_{ij}$$

The population variances of these three components are respectively $2\sigma_s^2$, σ_d^2 and σ_b^2 . When the data are analyzed, the deviation of the observed θ from their mean (designated $\bar{\theta}$) is easily separated into two parts, one a linear component in agreement with the law of comparative judgment, the other a residual component, as follows:

$$\underset{\text{total}}{(\theta_{ij} - \bar{\theta})} = \underset{\text{linear}}{(\theta^*_{ij} - \bar{\theta})} + \underset{\text{residual}}{(\theta_{ij} - \theta^*_{ij})}$$

Correspondingly we have the three sums of squares.

Total $$S_T \equiv \tfrac{1}{2} \sum_{i \neq j} (\theta_{ij} - \bar{\theta})^2$$

Linear $$S_L \equiv \tfrac{1}{2} \sum_{i \neq j} (\theta^*_{ij} - \bar{\theta})^2$$

Residual $$S_D \equiv \tfrac{1}{2} \sum_{i \neq j} (\theta_{ij} - \theta^*_{ij})^2$$

It may be noted that s , d , and b all affect the linear component (and also the total), while the residual is not affected by s , but only by d and b . This separation can now be used as the basis for an analysis of variance.

Because of the nature of the fitting process, and because of the slightly non-linear relation between angles and normal deviates, the deviations of the observations from their means have been separated into two parts which are not formally "orthogonal." There is no necessity for

$$\sum_{i \neq j} (\theta_{ij} - \theta^*_{ij})(\theta^*_{ij} - \bar{\theta})$$

to vanish. Consequently the two expressions for the sum of squares associated with the fit according to the law of comparative judgment,

$$\tfrac{1}{2} \sum_{i \neq j} (\theta^*_{ij} - \bar{\theta})^2 \equiv S_L$$

and

$$\tfrac{1}{2} \sum_{i \neq j} (\theta_{ij} - \bar{\theta})^2 - \tfrac{1}{2} \sum_{i \neq j} (\theta_{ij} - \theta^*_{ij})^2 \equiv S_T - S_D \quad ,$$

need not be precisely the same. So long as these give substantially the same answer, we may use either S_L or $S_T - S_D$ in assessing a "reliability" without serious error. (Should they differ widely, reconsideration of the fitting would be in order.)

The linear, residual, and total mean squares, together with the number of judges (N) and the number of stimuli (k), may be used

to give estimates of the variances as follows:

total mean square $$T \equiv \frac{2S_T}{k(k-1)} = \text{est } (2\sigma_s^2 + \sigma_d^2 + \sigma_b^2)$$

residual mean square $$D \equiv \frac{2S_D}{(k-1)(k-2)} = \text{est } (\sigma_d^2 + \sigma_b^2)$$

binomial mean square $$\frac{1}{N} = \text{est } \sigma_b^2$$

linear mean square $$L \equiv \frac{S_L}{k-1} = \text{est } (k\sigma_s^2 + \sigma_d^2 + \sigma_b^2)$$

It should be noted, as pointed out above, that we also have another possible value for the linear mean square given by

$$\frac{2(T-D)}{k} + D = \frac{S_T - S_D}{k-1} = \text{est } (k\sigma_s^2 + \sigma_d^2 + \sigma_b^2)$$

We may also define an associated set of chi-squares as follows:

$$\chi_T^2 = NS_T \quad , \qquad \chi_L^2 = NS_L \quad ,$$

$$\chi_{T-D}^2 = N(S_T - S_D) \quad , \qquad \chi_D^2 = NS_D \quad .$$

The basic formulas for the associated analyses are summarized in Table 4.

Starting with the observed values (p) and fitted values (p*) the values of θ and θ^* are found. These are used to compute S_T, S_D, and S_L, the sums of squares. From these we get the mean square values designated T, D, and L. These are used to give the estimates of variance components and "reliabilities."

The application of the procedure indicated in Table 4 to the data of Table 1 gives the results indicated in Table 5. In Table 5 the

TABLE 4

Outline of Analysis of Variance

Source of variation	Degrees of freedom (df)	Sum of squares	Mean square	Average value of mean square	Angular chi-square
All	$\frac{k(k-1)}{2}$	$\frac{1}{2}\sum_{i\neq j}(\theta_{ij} - \bar{\theta})^2 \equiv S_T$	$\frac{2S_T}{k(k-1)} \equiv T$	$2\sigma_s^2 + \sigma_d^2 + \sigma_b^2$	$\chi_T^2 = NS_T$
Law of comparative judgment (linear scale)	$k - 1$	$\frac{1}{2}\sum_{i\neq j}(\theta^*_{ij} - \bar{\theta})^2 \equiv S_L$ $S_T - S_D$	$\frac{S_L}{k-1} \equiv L$ $\frac{S_T - S_D}{k-1} = \frac{k(T-D)}{2} + D$	$k\sigma_s^2 + \sigma_d^2 + \sigma_b^2$	$\chi_L^2 = NS_L$ $\chi_{T-D}^2 = N(S_T - S_D)$
Residual, not accounted for by linear scale	$\frac{(k-1)(k-2)}{2}$	$\frac{1}{2}\sum_{i\neq j}(\theta_{ij} - \theta^*_{ij})^2 \equiv S_D$	$\frac{2S_D}{(k-1)(k-2)} \equiv D$	$\sigma_d^2 + \sigma_b^2$	$\chi_D^2 = NS_D$

"RELIABILITY" (r) OF SCALE

$$\rho_s = \frac{2\sigma_s^2}{2\sigma_s^2 + \sigma_d^2 + \sigma_b^2} \hat{=} \frac{2(L - D)}{kT} = r_s$$

$$\rho_s = \frac{2\sigma_s^2}{2\sigma_s^2 + \sigma_d^2 + \sigma_b^2} \hat{=} \frac{T - D}{T} = r_{ss}$$

$$\rho_b = \frac{2\sigma_s^2 + \sigma_d^2}{2\sigma_s^2 + \sigma_d^2 + \sigma_b^2} \hat{=} \frac{T - \frac{1}{N}}{T} = r_b$$

for two sets of scale values, x and y,

$$r_1 = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} \qquad r_2 = 1 - \frac{2\Sigma(x-y)^2}{\Sigma x^2 + \Sigma y^2}$$

VARIANCE COMPONENTS ANALYSIS

Source of variation	Symbol for variance component	Estimate of variance component
Binomial sampling	σ_b^2	$1/N$
Deviations from linear scale	σ_d^2	$D - 1/N$
Linear scale values of stimuli	σ_s^2	$\frac{L-D}{k}$ or $\frac{T-D}{2}$

$\theta_{ij} = 2 \arcsin \sqrt{p}$ (observed values)

$\theta^*_{ij} = 2 \arcsin \sqrt{p^*}$ (fitted values) (θ is in radians)

$\bar{\theta} = 2 \arcsin \sqrt{.5} = 1.571$

k = number of items, all pairs of which are compared.

N = number of judges for each pair.

$\hat{=}$ means here that average values are equal.

values obtained for the Chicago group are indicated by (C), the values for the Princeton group by (P), and the values obtained by pooling the numbers of judgments for the two groups are indicated by (T). The data on baseball teams presented by Mosteller (13) is indicated by (B).

The results show consistency in the variance components. Three estimates of the linear component are available in the handwriting experiment, 0.3521 (Chicago), 0.2868 (Princeton), and 0.3115 (combined). Three estimates are similarly available of a "deviations from scalability" component, 0.0166 (Chicago), 0.0171 (Princeton), and 0.0176 (combined). In comparison with the linear component the deviations components are small and agree unusually well among themselves. This fact suggests that we have systematic and consistent, though small, deviations from the law of comparative judgment.

Variance ratios.

In dealing with psychological tests many different sets of variance ratios have been used, giving various types of validity and reliability coefficients each having somewhat different properties and serving somewhat different purposes. In general these coefficients are the ratio of a measure of "true variance" to a measure of "observed variance" which includes both "true variance and error variance." One reasonable interpretation for paired comparisons is to regard the linear component ($2\sigma_s^2$) as "true variance" and the other two components ($\sigma_d^2 + \sigma_b^2$) as error variance, so that we may define a coefficient of "linear consistency" through

$$\rho_s = \frac{2\sigma_s^2}{2\sigma_s^2 + \sigma_d^2 + \sigma_b^2} \doteq \frac{2(L - D)}{kT} = r_s$$

TABLE 5

Comparison of Scaling Data

Analysis of Variance

Source of variation	Degrees of freedom (df)	Sum of squares	Mean square		Angular chi-square	P
		S_T	T		χ^2_T	
All	36	26.7717	.7437	(C)	2677.17	(<.00001)
		21.7404	.6039	(P)	2174.04	
		23.6709	.6575	(T)	4734.18	
	28	3.3468	.1195	(B)	73.63	(<.0001)
		S_L	L	$\frac{S_T - S_D}{k - 1}$	$\chi^2_L = NS_L$	
Linear scale	8	25.5606	3.1951	(3.2534) (C)	2556.06	(<.00001)
		20.8661	2.6083	(2.6229) (P)	2086.61	
		22.6075	2.8259	(2.8796) (T)	4521.50	
	7	2.6813	.3830	(0.3822) (B)	58.99	(<.0001)
		S_D	D		$\chi^2_D = NS_D$	
Residual	28	.7449	.0266	(C)	74.49	(<.0001)
		.7575	.0271	(P)	75.75	
		.6338	.0226	(T)	126.76	
	21	.6717	.0320	(B)	14.78	(.80+)

Estimated Variance Components

	Linear scale values, σ^2_s		Deviations from scalability, σ^2_d	Binomial variation, σ^2_b
	$\frac{L-D}{k}$	$\frac{T-D}{2}$		
(C)	.3521	.3585	.0166	.0100
(P)	.2868	.2884	.0171	.0100
(T)	.3115	.3174	.0176	.0050
(B)	.0439	.0437	-.0135	.0455

Estimated Reliabilities

	r_s	r_{ss}	$r^2_{\theta\theta*}$	r_b
	$\frac{2(L-D)}{kT}$	$1 - \frac{D}{T}$		$1 - \frac{1}{NT}$
(C)	.9468	.9642	.9723	.9866
(P)	.9498	.9551	.9652	.9834
(T)	.9475	.9656	.9733	.9924
	(r_2 = .956)			(r_1 = .989)
(B)	.7343	.7322	.7993	.6192

	k	N
(C) = Chicago data	9	100
(P) = Princeton data	9	100
(T) = These two together	9	200
(B) = Baseball data from Mosteller (13)	8	22

The factor 2 arises from the fact that σ_s^2 was normalized in terms of individual stimuli, while σ_d^2 and σ_b^2 are normalized in terms of differences. That is, σ_s^2 is the variance of the k different s -values, while the variance of the $k(k - 1)$ values $(s_i - s_j)$ is $2\sigma_s^2$, and the observed variance for the cell entries is $(2\sigma_s^2 + \sigma_d^2 + \sigma_b^2)$.

If the linear sum of squares is taken as $S_T - S_D$ (instead of S_L), then we have another estimate for the coefficient of linear consistency.

$$r_{ss} = \frac{T - D}{T} \hat{=} \frac{2\sigma_s^2}{2\sigma_s^2 + \sigma_d^2 + \sigma_b^2} = \rho_s$$

These coefficients r_s and r_{ss} indicate the extent to which the linear model (as represented by the fitted values θ^*) fits the observed cell entries, given by θ . For example, if the agreement is perfect, then S_D and D will equal zero, S_T will equal S_L which means that $2L/k = T$ so that $r_s = r_{ss} = 1.00$. If, on the other hand, the mean squares T , L , and D are all equal, then $r_s = r_{ss} = 0.00$. These coefficients r_s and r_{ss} are regarded as similar to $r^2_{\theta\theta^*}$, the square of the correlation between observed and true values assuming the linear model. Alternatively, r_s and r_{ss} may be regarded as representing the correlation between two sets of observed values provided their correlation is entirely accounted for by the true values (assuming a linear model). The coefficients r_s or r_{ss} may be regarded as appropriate to the recomparison of a randomly selected pair of the nine handwriting specimens against a background of seven other specimens

covering the same range of merit and hence drawn from a population having the same σ_s^2 as the specimens used in this experiment. For example, if another set of three specimens each of values 50, 70, and 80 were scaled, a similar σ_s^2 would be expected; if σ_d^2 and σ_b^2 also remained about the same, a similar degree of agreement between fitted and observed values, i.e., a similar coefficient of linear consistency, would be expected.

However, if all the handwriting specimens (from 10 to 90) in the Ayres Scale were used, one would expect a larger σ_s^2, and if, as seems plausible, σ_d^2 remained about the same, the result would be a higher coefficient than that found here using only values 50, 70, and 80. On the other hand, if one used only specimens 50, 60, and 70, a slightly smaller σ_s^2 and (if σ_d^2 remained about the same) a slightly lower reliability would be expected.

It can be seen that even though Mosteller's chi-square goodness of fit test (χ_D^2) shows clearly that the handwriting data deviates significantly from a linear scale, nevertheless the scales show a satisfactory agreement with the linear model, about .95 for the case where the nine handwriting specimens were rated by 100 or 200 judges. Since only $2\sigma_s^2$ is considered to be true variance, the coefficients given by r_s and r_{ss} will be what are usually termed "conservative" estimates. A "dashing" estimate for reliability is obtained by regarding σ_d^2 as part of the true variance rather than as part of the error variance. Thus we have

$$\rho_b = \frac{2\sigma_s^2 + \sigma_d^2}{2\sigma_s^2 + \sigma_d^2 + \sigma_b^2} \hat{=} \frac{T - \frac{1}{N}}{T} = r_b$$

This definition yields for the handwriting data reliabilities of .98 or .99. This coefficient represents the correlation between two sets of θ -values for the same stimuli judged by another random sample of people from the same population. Coefficients computed from this formula are appropriate to the recomparison of a randomly selected pair of the nine specimens against a background of seven other handwriting specimens drawn from a population having the same σ_s^2 and also the same peculiarities that produced the deviations from linearity. One possibility is a recomparison of a random pair against a background of the same seven other handwriting specimens. Thus we see that without any assumptions about the law of comparative judgment one has a set of stimuli that cannot be regarded as indifferent to the subjects.

A corresponding chi-square is given by

$$\chi_T^2 = NS_T$$

with degrees of freedom

$$df = (k/2)(k - 1)$$

These values of chi-square (χ_T^2 in Table 5) are all extremely large, indicating a negligible probability that the data could have arisen by random sampling from a population in which the proportions were all .5.

The coefficient (r_b), which is zero if the percentages of Table 1 are all random binomial deviations from .5, may be compared with Kendall's coefficient of agreement (10, pp. 125ff.; 11, pp. 333ff.),

which is unity only if all proportions are 1.0 or 0.0; i.e., if there is complete agreement among all judges in making each judgment. Kendall's coefficient of agreement is determined directly from the experimental frequencies, without using any transforms such as the arc sin. The data here presented cannot be regarded as showing such agreement among all judges. However, it clearly cannot be regarded as indicating only random judgments.

We may compare these coefficients computable for a single set of data with more conventional reliabilities obtained by comparing the Princeton with the Chicago scale values. The correlation between the two sets of values in Table 2 (r_1) is .989, which, it may be noted, is similar in magnitude to r_b . If we make no allowance for changes in discriminal dispersion, but take the entire difference of scale values (adjusted to a common mean but not to a common variance) as error, then

$$r_2 = 1 - \frac{2\Sigma(x - y)^2}{\Sigma x^2 + \Sigma y^2} = .956$$

which is similar in magnitude to the estimates of ρ_s .

Two coefficients have been suggested. The coefficient r_b indicates the extent to which the stimuli are differentiated by the subjects.

It seems reasonable to regard r_s or r_{ss} as a conservative estimate of consistency for a single set of data scaled by the law of comparative judgment. In such a case there would be no replication to indicate that σ_d^2 might, from some points of view, reasonably be regarded as part of the true variance. The estimates r_s and r_{ss} give

a direct measure of the agreement between the observed (θ) and fitted (θ^*) values of the arc sin $\sqrt{p}$.

The lines labelled "(B)" in Table 5 give for comparison the data on baseball teams reported by Mosteller (13). It is interesting to note that despite the non-significant chi-square, the reliability (r_s or r_{ss}) is only .73, while r_b = .62 . This low reliability is due apparently to the similarity of the different teams, since est σ_s^2 is only .0439, which is less than the binomial variation of .0455 with which σ_s^2 must be combined. Under these circumstances it is not surprising that chi-square is not significant, especially with N as low as 22. On the other hand, the data on handwriting has a smaller binomial variance (.01), and a much larger σ_s^2 (about .3). Despite the fact that the residual mean square (D) is slightly smaller than that for the baseball data, when N equals 100 or 200 with 28 degrees of freedom, this much smaller discrepancy cannot be regarded as due to chance.

In summary, a variance-components analysis has been presented for paired comparisons. This analysis gives estimates of the variance of the actual scale values (σ_s^2), and the variance of observations due to deviations of the data from the linear paired comparisons model (σ_d^2), which are compared with the binomial sampling variance (σ_b^2). A variety of coefficients based on these three variances are also presented. If one is interested in asking whether or not the subjects' responses are purely random, then Kendall's coefficient of agreement, or the r_b as presented here may be used. If one is interested in the extent to which the law of comparative judgment accounts for the data, then r_s or r_{ss} would be the appropriate coefficient.

BIBLIOGRAPHY

1. Ayres, L. P. A scale for measuring the quality of handwriting of school children. New York: Russell Sage Foundation, 1912, Pp. 90.

2. Bennett, Carl A., and Franklin, Norman L. Statistical analysis in chemistry and the chemical industry. New York: John Wiley and Sons, 1954.

3. Cochran, W. G. The X^2 test of goodness of fit. Ann. math. Stat., 1952, 23, 315-345.

4. Davies, Owen L. (Ed.) Design and analysis of industrial experiments. London: Oliver and Boyd, and New York: Hafner Publishing Co., 1954.

5. Duncan, A. J. Quality control and industrial statistics. Chicago: Richard D. Irwin, Inc., 1952, Pp. xxvii + 663.

6. Fisher, R. A. Statistical methods for research workers. (10th Ed.) London: Oliver and Boyd, 1946.

7. Fisher, R. A. and Yates, F. Statistical tables for biological, agricultural and medical research. London: Oliver and Boyd, 1938, and New York: Hafner Publishing Co., 1953, Pp. 90.

8. Freeman, M. F. and Tukey, J. W. Transformations related to the angular and the square root. Ann. math. Stat., 1950, 21, 607-611.

9. Hald, A. Statistical tables and formulas. New York: John Wiley and Sons, 1952.

10. Kendall, M. G. Rank correlation methods. London: Chas. Griffin and Co., 1948.

11. Kendall, M. G. and Babington Smith, B. On the method of paired comparisons. Biometrika, 1940, 31, 324-345.

12. Mood, A. M. Introduction to the theory of statistics. New York: McGraw-Hill, 1950, Pp. 433.

13. Mosteller, Frederick. Remarks on the method of paired comparisons. III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed. Psychometrika, 1951, 16, 207-218.

14. Thurstone, L. L. Psychophysical analysis. Amer. J. Psychol., 1927, 38, 368-389.

15. Thurstone, L. L. A law of comparative judgment. Psychol. Rev., 1927, 34, 273-286.

16. Tippett, L. H. C. The methods of statistics. (4th Ed.) New York: John Wiley and Sons, 1952, Pp. 395.